Supplementary Material

AlleleSeq: Analysis of Allele-Specific Expression and Binding in a Network Framework

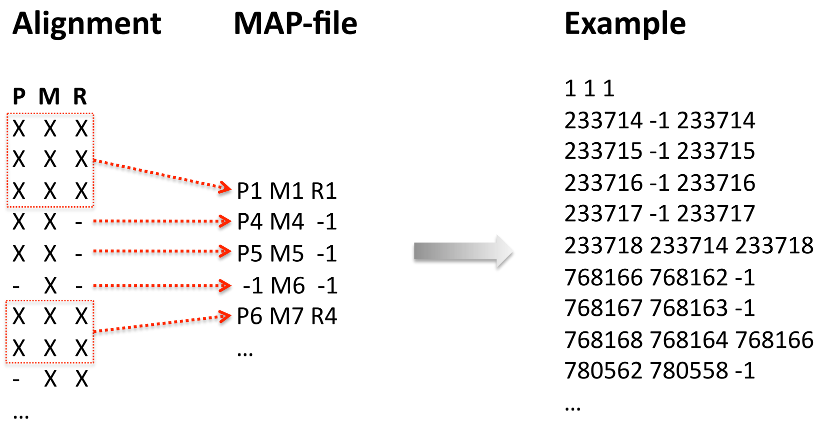
Joel Rozowsky[*,†,1,2], Alexej Abyzov[*,1,2], Jing Wang[2], Pedro Alves[2], Debasish Raha[3], Arif Harmanci[1,2], Jing Leng[2], Robert Bjornson[4,5], Yong Kong[5], Naoki Kitabayashi[6], Nitin Bhardwaj[1,2], Mark Rubin[6], Michael Snyder[7] and Mark Gerstein[†,1,2,4].

* These authors contributed equally.

† Correspondence should be sent to pi@gersteinlab.org

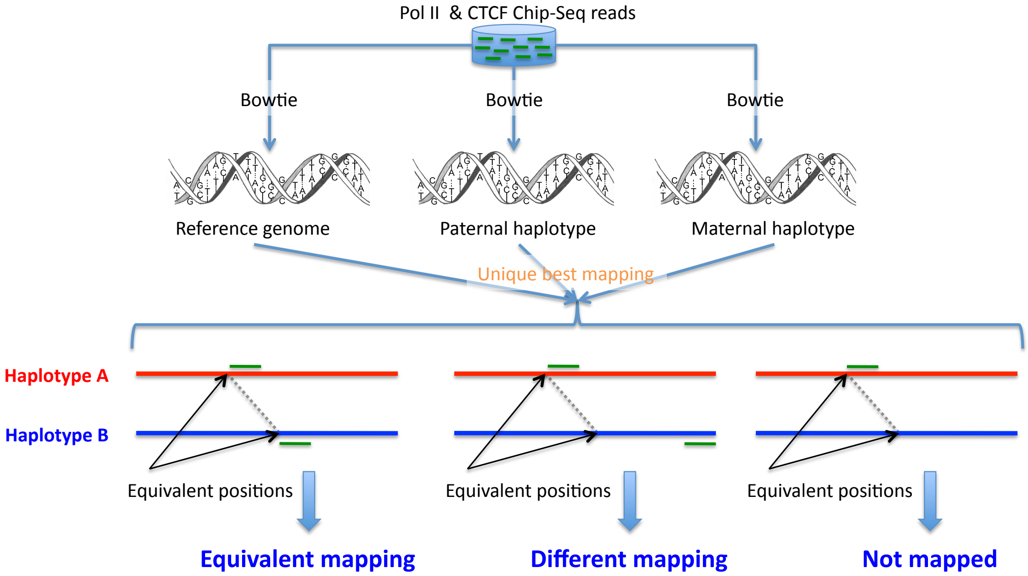
Table of Contents

[*] Also available from http://alleleseq.gersteinlab.org/

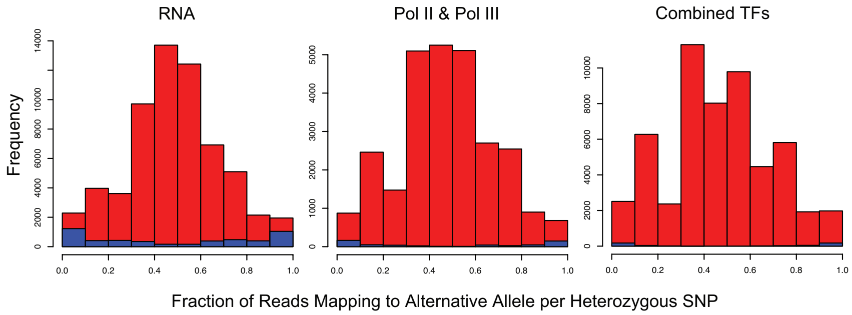


```
1 1 1
233714 -1 233714
233715 -1 233715
233716 -1 233716
233717 -1 233717
233718 233714 233718
768166 768162 -1
768167 768163 -1
768168 768164 768166
780562 780558 -1
...
```

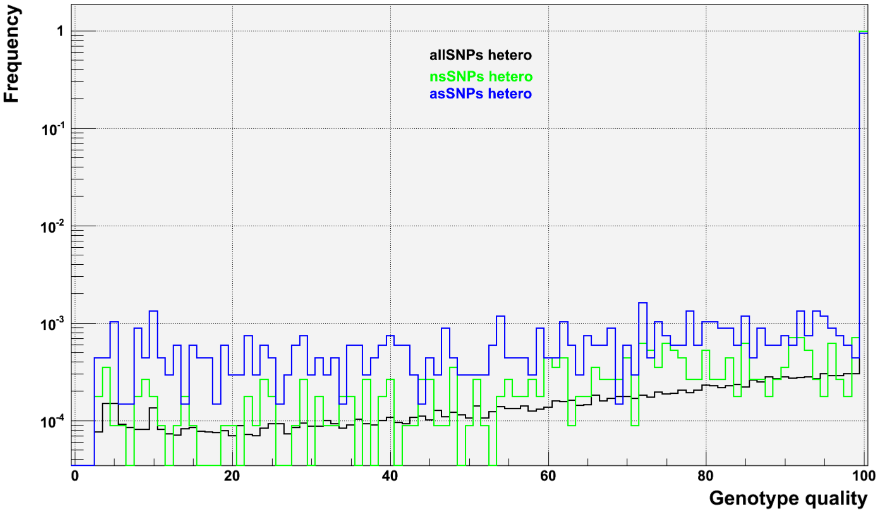
Supplementary Figure 1. MAP file establishes equivalence of bases between reference haplotypes of personal genome and reference genome. Paternal haplotypes denoted as R, maternal as M, and reference as R. Nucleotides in each sequence are denoted as X. Ungapped block that is present in all three haplotypes is recorded by indices of the first bases in a block for each haplotype. Each position absent in either haplotype is recorded by base indices for haplotype(s) having a nucleotide in the position and by -1 for haplotype(s) with gap. Base numbering starts from 1.



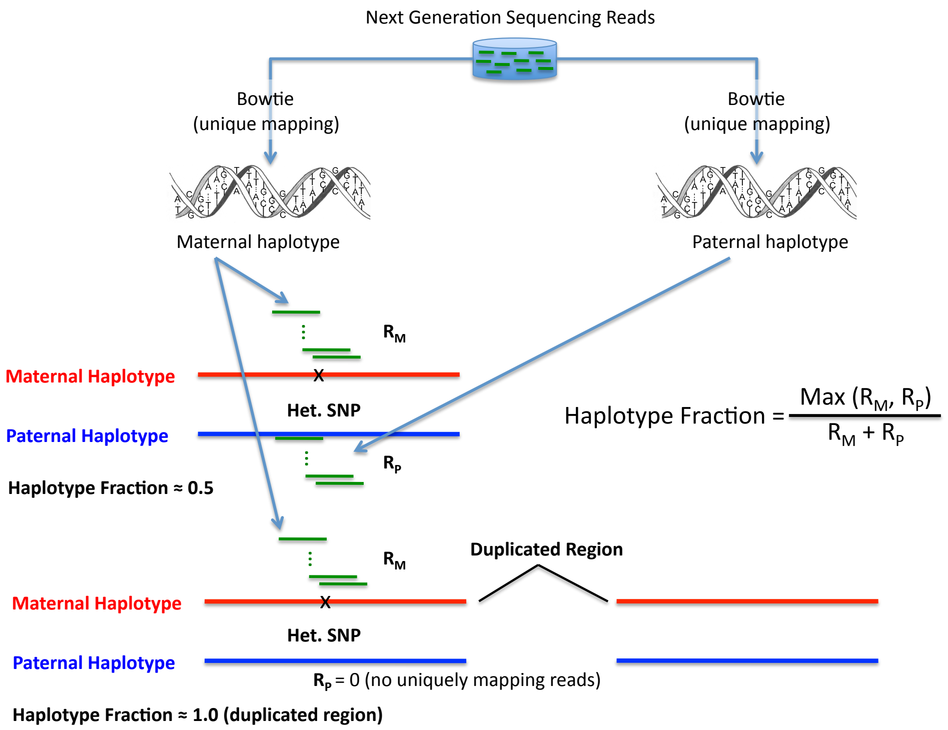
Supplementary Figure 2. Procedure to compare read mapping to reference genome and haplotypes of NA12878. Pol II Chip-Seq reads were mapped to each haplotype using Bowtie. Then, for any pair of compared haplotypes read mappings were classified as i) equivalent, if a read maps at the equivalent positions; ii) different, if a read maps to non-equivalent positions; iii) not mapped, if a read does not map to either of the compared haplotypes.



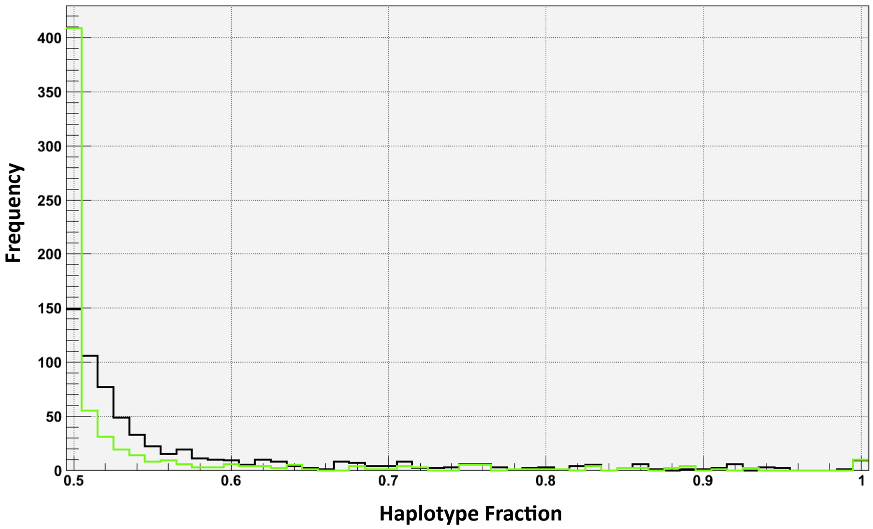
Supplementary Figure 3. For each heterozygous SNP location covered at a depth greater than six we can compute the fraction of reads derived from the maternal allele relative to the paternal sequence. We then have plotted the distribution of maternal allele fraction for all heterozygous SNPs (significant allele-specific positions are indicated in blue) for the RNA-Seq, Pol II and remaining ChIP-Seq datasets combined. We observe that the distribution of all heterozygous SNPs as well as the allele-specific SNP positions are both quite symmetric and thus we do not see a significant bias towards either the maternal or paternal haplotype.



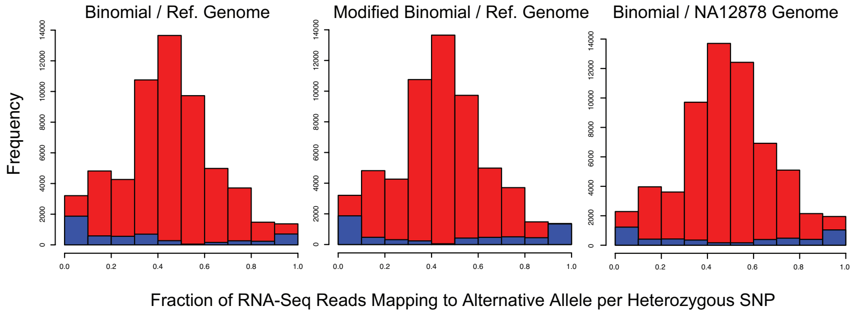
Supplementary Figure 4. Density distribution of heterozygous SNPs in NA12878 by genotyping qualities (all SNPs -- black line, SNPs that are identified as allele specific in this study -- blue line, non-synonymous SNPs – green line). AS SNPs are on average of slightly lower quality than all SNPs, however the vast majority of them (~99%) are confidently genotyped (quality score > 30). Non-synonymous SNPs also exhibit a very similar distibution.



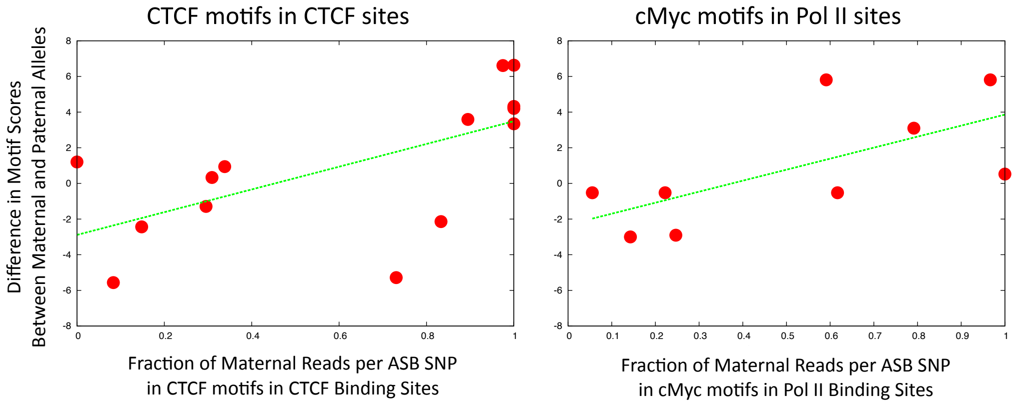
Supplementary Figure 5. In this schematic we show the potential effect of a genome duplication on a region showing allele-specific behavior (as described in Degner *et al.* 2009). In a unduplicated regions if we align the reads uniquely to each haplotype independently we can compute the haplotype fraction (the fraction of reads mapping to one haplotype over the reads mapping to both haplotypes) overlapping a heterozygous SNP locations exhibiting allele-specific behavior. For unduplicated regions this fraction should be approximately 0.5. However, for genomic regions that are duplicated (the heterozygous SNP is only at one of the locations) then we would observe allele-specific behavior since no reads would align uniquely to one allele (the paternal haplotype in this example). However, the haplotype fraction would necessarily by close to 1. We can use this fraction to determine the number of allele-specific sites that are caused by genome duplications.



Supplementary Figure 6. Read mapping skew towards either haplotype at ASB site for Pol II. For each ASB site the skew was calculated as a fraction of reads mapped to one haplotype over the sum of reads mapped to each haplotype. At each site the maximum value of skew is used. Black line shows the distribution of skew when reads mapping uniquely to each either haplotype are considered. Green line shows the distribution of skew when read mapping to each haplotypes are used. Sites with large skew (> 0.6) can be the result of read mapping bias suggested earlier (Degner *et al.* 2009), when reads on one haplotype couldn't be mapped due to non-unique mapping. However, the skew is still observed for a number of sites when the reads that map to only one haplotype are excluded from consideration (green line) to prevent mapping bias. To be conservative we considered an ASB site not affected by mapping bias if mapping skew does not change by more than 0.1 if all reads or only reads mapping to each haplotype are considered. This estimates that less than 15% of ASB sites can be affected by read mapping bias. Note, however, that alternative explanation is that those sites have other variants within the read length, and due to that read coming from one haplotype do not map to another.



Supplementary Figure 7. For each heterozygous SNP location covered at a depth greater than six we can compute the fraction of reads derived from the alternative allele relative to the reference sequence. We then plotted the distribution of alternative allele fraction for all heterozygous SNPs (significant allele-specific positions are indicated in blue) for the RNA-Seq data. The left panel show the distribution using a unmodified binomial distribution and using reads aligned to the reference genome, the middle panel shows the results using a modified binomial distribution (Montgomery *et al.* 2010) using reads aligned to the reference genome and the right shows the distribution using reads aligned against the diploid genome for NA12878 (the same as Figure 2) using an unmodified binomial distribution. We observe that the most naïve approach (left panel) is significantly skewed towards the reference allele, while using the modified binomial approach the significant SNPs are more symmetric however not as symmetric as the approach we use in this paper.



Supplementary Figure 8. We plot the difference of motifs scores (see **Methods**) between the maternal and paternal alleles against the fraction of maternally derived reads for ASB SNPs overlapping motifs within binding sites. On the left we plot this for ASB SNPs in CTCF motifs that are located within CTCF binding sites. On the right we plot this for ASB SNPs in cMyc II motifs that are located within Pol II binding sites.

Coverage Threshold	Total number of heterozygous SNPs detected	Number of 1000 genomes heterozygous SNPs detectable	Number of 1000 genomes heterozygous SNPs detected
10x	69,289	47,598	17,701
25x	51,584	20,598	15,097
50x	29,752	11,234	8,528
100x	14,706	5,663	4,224

Supplementary Table 1. In this table we show the results of determining heterozygous SNPs denovo from the RNA-Seq data using SNVmix (Shah *et al.* 2009). Requiring different coverage by RNA-Seq reads (i.e. expression levels) we observe the number of heterozygous SNPs detected genome wide in the first column. In the second column we show the number of 1000 genomes heterozygous SNPs for NA12878 detectable at each coverage threshold and in the third column the number of these that are detected. We observe that even with a high coverage threshold a substantial number of false positive heterozygous SNPs are detected by denovo SNP calling on the functional genome reads.

Haplotype	# of mapped reads	Equivalently mapped reads in		
		Reference	Paternal	Maternal
Reference	26,322,823		26,287,466 (99.87%)	26,311,193 (99.96%)
Paternal	(+0.24%) 26,386,899	26,287,466 (99.62%)		26,334,565 (99.80%)
Maternal	(+0.33%) 26,411,779	26,311,193 (99.62%)	26,334,565 (99.71%)	
		Differently mapped reads in		
		Reference	Paternal	Maternal
Reference	26,322,823		6,579 (0.02%)	7,013 (0.03%)
Paternal	(+0.24%) 26,386,899	6,579 (0.02%)		31,134 (0.12%)
Maternal	(+0.33%) 26,411,779	7,013 (0.03%)	31,134 (0.12%)	
		Unmapped reads in		
		Reference	Paternal	Maternal
Reference	26,322,823		28,778 (0.11%)	4,617 (0.02%)
Paternal	(+0.24%) 26,386,899	92,854 (0.35%)		21,200 (0.08%)
Maternal	(+0.33%) 26,411,779	93,573 (0.35%)	46,080 (0.17%)	

Supplementary Table 2. Comparison of read mappings to reference genome and paternal and maternal haplotypes of GM12878 (similar to Table 3). Chip-Seq reads for CTCF were independently mapped to each haplotype (chromosomes 1-22 and X) and the best unambiguous mapping (no more than two mismatches) was selected for each read.

Pol II Overlap	Maternal	Paternal	Reference
Maternal	1.000	0.978	0.957
Paternal	0.966	1.000	0.949
Reference	0.956	0.960	1.000
CTCF Overlap	Maternal	Paternal	Reference
Maternal	1.000	0.992	0.985
Paternal	0.991	1.000	0.985
Reference	0.986	0.987	1.000

Supplementary Table 3. In this we independently mapped reads for both Pol II and CTCF ChIP-Seq againt the maternal and paternal haplotypes as well as the reference (hg18/NCBI36) genome sequences. Using PeakSeq (Rozowsky et al. 2009) we determine binding sites for each of these three genomes using the same parameters. We then perform a pair-wise nucleotide overlap of the binding sites between the three-genome sequences for both Pol II and CTCF. We observe that in both cases the overlap is better between the maternal and paternal genomes than compared against the reference sequence. An additional observation is that the difference between binding sites for the three genomes is greater than between the reads used (see Table 3 and Supplementary Table 2.

ASE and ASB from Heterozygous Indels	Total Count	Maternal	Paternal
Exons showing ASE	128	75	53
Novel TARs showing ASE	233	126	107
Pol II binding sites showing ASB	123	53	70
CTCF binding sites showing ASB	52	22	30

Supplementary Table 4. In this table we present the number of additional sites exhibiting allele-specific behavior for RNA-Seq known exons and novel TARs determined by an exon or novel TAR overlapping a heterozygous indel. We also show the number of additional ASB binding sites for Pol II and CTCF for detemined for binding sites overlapping a heterozygous indel.